

Big Data Science and Civil Infrastructure

Dr. Alok Aggarwal

CEO and Chief Data Scientist

October 09, 2015

Scry Analytics





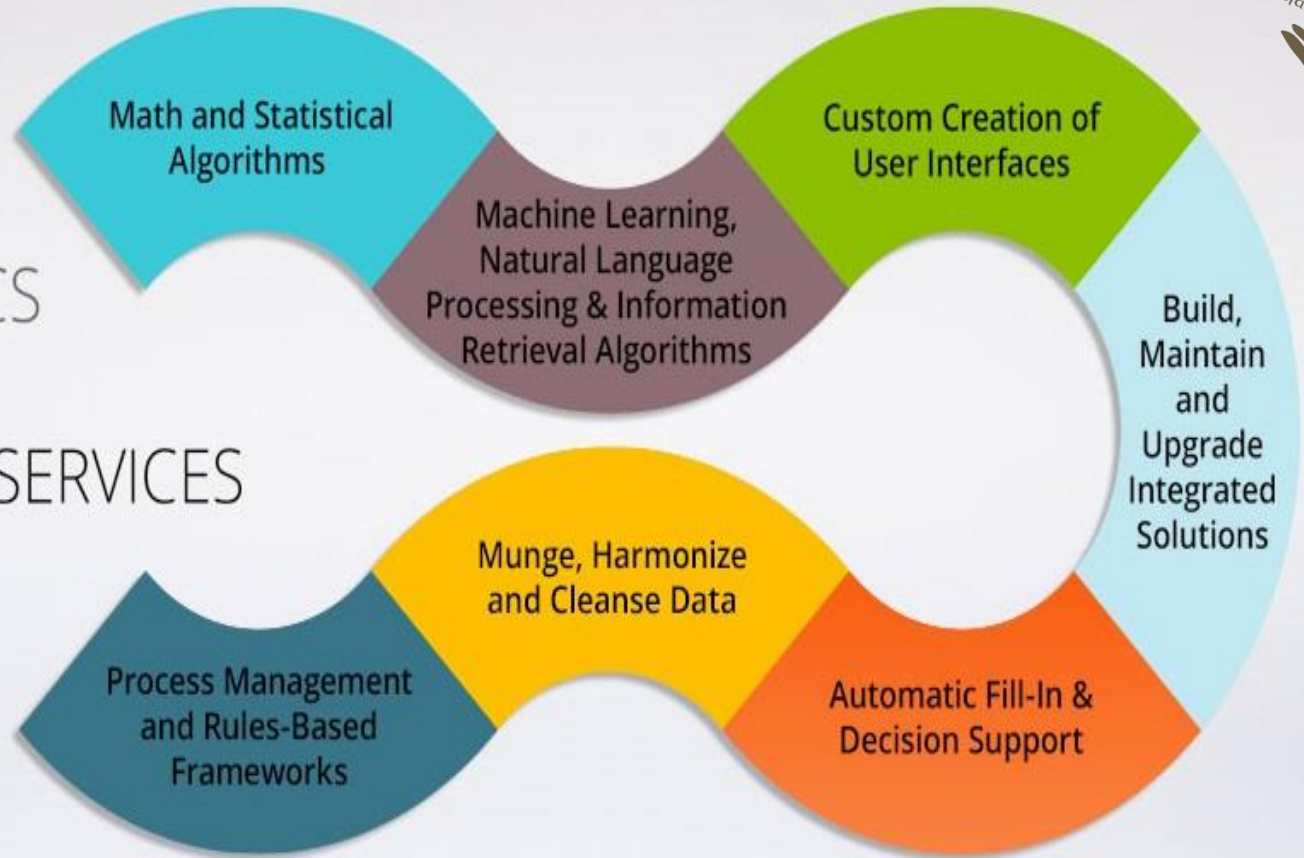
About Scry Analytics

Founded in Feb - 2014, Scry Analytics (www.scryanalytics.com) is using automation and artificial intelligence for the services industry to improve processes and work-flows with respect to Profitability, Risk, Experience, Compliance, Quality & Timeliness (PRECQT)

- Our computation platform uses open source software and proprietary algorithms
- Our domain experts and math & computer scientists use various kinds of data (e.g., structured, unstructured, machine logs) to provide solutions and services
- Provide real-time or near real-time analysis
- Underlying algorithms include e.g., classical analytics, machine learning, natural language processing and speech analysis algorithms

Scry Analytics combines subject matter expertise with automated computation and artificial intelligence to enhance data driven decision making under uncertainty

SCRY ANALYTICS
INNOVATIVE
SOLUTIONS & SERVICES



LOCATIONS

San Jose, California; Raleigh, North Carolina; Delhi-Gurgaon, India

What is Big Data?



- No precise definition
- 1997: Michael Cox and David Ellsworth from NASA published “[Application-controlled demand paging for out-of-core visualization](#)” in Proc. of the IEEE 8th Conf. on Visualization -- “Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.”
- 2001: Doug Laney from MetaGroup (now Gartner) provided “Velocity, Variety & Volume” as three vital attributes of Big Data and since then others have added “Value, Veracity & Variability”

Data Warehouses and Data Lakes



Limitations of Data Warehouses:

- Store data from various sources in specific static structures, aggregates, and categories
- Usually, only predefined analysis can be done on these structures and the scope is restricted to the canned reports, dashboards (with limited and pre-defined modes of interaction)
- Hard to include new data sources later; e.g., machine data, sensor data, weather and climate, soil, highway-bridge related data, rules and regulations related (i.e., natural language) data
- Falls apart in the world of big data where after we have a set of models that are working well with the current data sets, we would like to improve them by bringing in newer and different (and often unstructured) data and algorithms, e.g., social media, web data or machine data

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Advances in Technology During 1992-2015



**1965 – 2015
Moore's Law**

- Gordon Moore's Law: For any given price, the processing power (& memory) doubles every 18 months.
- Hence, a program written in 1992 be now run in three minutes instead of 3,000 minutes (or two days).
- Computational Power and Memory is now a commodity (available on Amazon or Azure)

**2004 – MapReduce
2006 – Apache Hadoop**

- Google introduced MapReduce for parallel & distributed computing of large data.
- Using MapReduce, Google solved a single problem using 40,000 computers.
- Yahoo! modified Google's MapReduce to create a programming language called Apache Hadoop. Since the product is free, it has made Hadoop extremely popular.

**Improved Algorithms
(e.g., Deep Learning Algorithms)**

- Since computational power and memory is now abundantly available, theoretical research in parallel and distributed computing is now being implemented and practiced.
- New research regarding Deep Learning and Error Correcting Codes is already beginning to show results (both in lab settings and for real live problems).

Key Performance Indicators for Most Work-flows (T, Q, C, E & R)

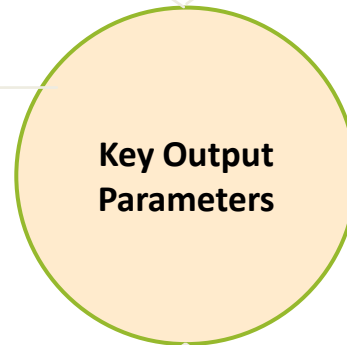


By and large, firms are interested in getting insights for improving the following metrics for their work-flow systems:

T: Improving **Timeliness** in providing their products, services and solutions

Q: Improving **Quality** of their product, services, solutions and that of their work-flow systems

C: Increasing **cash flow**, profit and revenue



E: Enhancing customer **Experience** to ensure improved customer loyalty and brand

R: Reducing **Risks** with respect to operations and market conditions and ensuring Compliance with respect to rules, regulations and laws

External Data Often Required to Improve T, Q, C, E and R

To improve a Work-flow's Timeliness, Quality, Cash-flow, Experience and Risk, a business often needs to **Analyze and Gain Insight** by including the following disparate data:

Market research data, syndicated research data, demographic data, economic data, internal and external transaction data, weather data, call center data, machine data, social media data, and ERP & CRM data, etc.

Improving Performance Indicators in Railroads



Big Data Science can be used for:

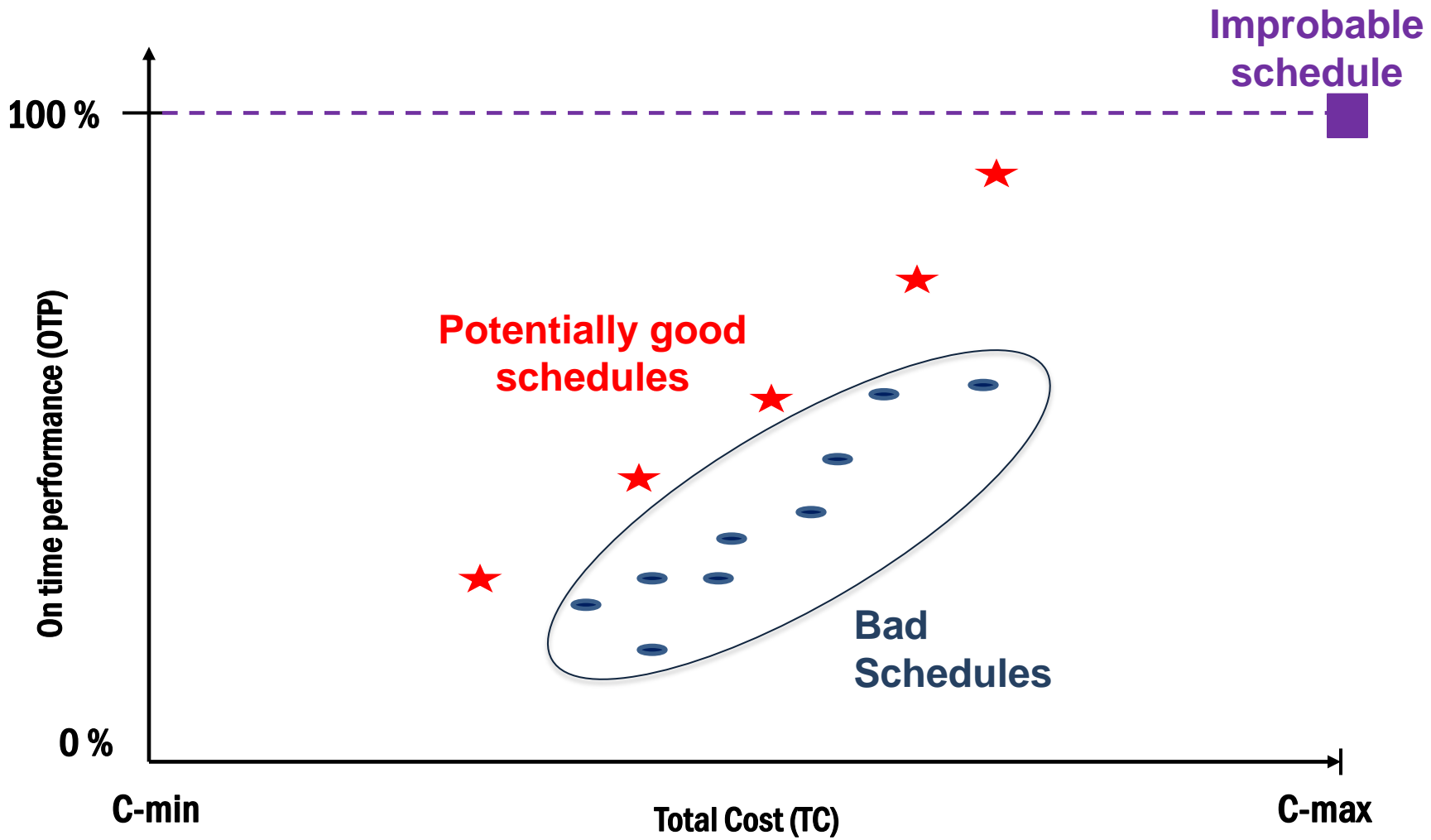
- Predicting Equipment failure
- Sensors data analysis
- Preventive maintenance
- Maintenance Planning
- Financial Planning



By doing advanced analytics and machine learning and by incorporating the following data, we can predict how many wheels are going to fail next year (by month, car type, size, location, etc.) with 2%:

- New cars
- Expected new mileage
- Macro-economic conditions (and Micro-economic also)
- Daily weather conditions
- Seasons, holidays, etc.
- Freight and/or Loading
- Attributes of the wheel
- Data from sensors (and determining, which if any sensors are faulty)

Multi-Variable Optimization for Railroads



Four Pillars of Big Data Science



Cleanse and Harmonize Data:

- Different kinds of data is coming in from Internal and External sources
- Depending upon the problem, 10%-40% work is spent in data cleansing
- Need to build a Thesaurus for unstructured data (i.e., words) for a specific domain
- At a minimum, ensure that all disparate data is connected via key words, key indexes etc.
 - Goal: > 80% cleansing by software; < 20% by professionals
 - On-going Process: Cleanse new & different data

Provide a Holistic Data Science Solution For A Given Problem

Subject Matter Expertise:

- Same words may be different in different verticals, e.g., IP is Internet Protocol in Tech & Telecom industry but is Intellectual Property in Law
- Professionals within the same group may not agree on the meaning of the same term

Math and Computer Science:

- Plethora of math & comp. science algorithms and models already exist in open source (or licensed) software; some need to be proprietary
- Key is to iterate and improve known algorithms that work best for the given problem

Obtain Raw Data:

- Current data may lie in different databases, data warehouses, emails, or other documents
- May lie in different business Units, different countries, and in different formats (e.g., structured, unstructured)

Big Data for Improving Bridges



Large bridges have 1,500 + sensors of various forms; numbers will only increase; travelers' feedback will need to be incorporated also

Stonecutters Bridge (Ni & Wong 2012)

- Accelerometers (58)
- Temperature Sensors (388)
- Dynamic Strain Gauges (678)
- Static Strain Gauges (158)
- Global Positioning Systems (20)
- Displacement Transducers (34)
- Bearing Sensors (12)
- Barometers/Rainfall Gauges/Hygrometers (28)
- Corrosion Cells (33)
- Digital Cameras (18)
- Weigh-in-Motion Stations (4)

Total No. of Sensors : 1505



Stonecutters Bridge , Hong Kong

Volume, Velocity, Value:

- 145 sensors sampling at 100Hz produce ~56 kB/sec; 5GB/day; 1.75 TB per year;
- Each video camera produces 46kB/s; 1.3 TB/year
- Some computation should be done at edge of network

Big Data for Improving Bridges



Variety, Veracity & Variability:

- Different types of sensors will increase; Internet of Things
- Sampled at different intervals
- Different actions and varying demands and responses
- Different accuracy of data (from one sensor to another)
- Legacy systems' data will be hard to munge & harmonize
- Missing data
- External data (e.g., weather and traffic patterns)
- Additional data to check if sensors are malfunctioning
- Effect of temperature, pressure, wind etc. may influence measurements
- Feedback data from travelers
- Data from test results of various materials



Sheikh Rashid bin Saeed Crossing, Dubai

- Data (structured and unstructured data) from human “visual” inspections
- Data may stored as drawings and/or quantitative & qualitative reports

Thank you

For more information, please contact:

alok.aggarwal@scryanalytics.com

+1 408 872 1078 T

+1 914 980 4717 M

Scry Analytics

